Submission in Response to NSF CI 2030 Request for Information

DATE AND TIME: 2017-04-05 16:26:21
REFERENCE NO: 285

Author Names & Affiliations

- Gavin Conant - Department of Biological Sciences, Bioinformatics Research Center, Program in Genetics - North Carolina State University
- Michela Becchi - Department of Electrical and Computer Engineering, North Carolina State University

Contact Email Address (for NSF use only)

(Hidden)

Research Domain, discipline, and sub-discipline

Biology, genomics, computational genomics

Title of Submission

Extensible, high-performance cyberinfrastructure for enabling biological discovery from extreme-scale sequence datasets

Abstract (maximum ~200 words).

The per-megabase cost of DNA sequencing has fallen so rapidly that problems in biology that were long thought intractable, such as genetically cataloging entire ecosystems or complete surveys of genetic variation across a population, are becoming feasible. Unfortunately, the computational tools for employing these vast datasets have not kept pace with the growth in sequencing capacity. As such, there is a need for new computational tools that take advantage of recent advances in computing (such as accelerator computing and cloud-based services) in a way that allows non-expert users to employ the full power of modern parallel architectures. We illustrate this need for new algorithms and tools with three applications: population genomics, comparative genomics, and metagenomics.

Question 1 Research Challenge(s) (maximum ~1200 words): Describe current or emerging science or engineering research challenge(s), providing context in terms of recent research activities and standing questions in the field.

With the advent of massively parallel DNA sequencing technologies (Shendure and Ji 2008), molecular biologists have begun to experience directly the challenges inherent in very-large scale datasets. The field was arguably late to the party: physics, astronomy, and climate science had had to learn to accommodate large data volumes at least a decade earlier (Stoughton et al. 2002; Lamanna 2004; Meehl et al. 2007). The particle physics example is a useful one: the computing design of particle accelerators like the Large Hadron Collider were explicitly designed with the expectation of very high rates of data arrival and comprise both permanent primary data storage and near-real time data reduction to allow scientific discovery on more manageable dataset sizes (Lamanna 2004). Hence, we argue that biologists should resist the temptation to manage these increasingly large datasets by seeking to trivially “scale-up” existing codes and algorithms. Instead, the size of next-generation sequence datasets needs to be seen as a facet of a larger trend in science: the increasing

complexity of the data, models and software needed for scientific discovery.

Computing is the key tool for managing these large data volumes (Schadt et al. 2010) and the problems posed by large datasets are compounded by the belated arrival of the long-predicted demise of Moore's Law for single-threaded CPU performance (National Research Council (U.S.). Policy and Global Affairs. 2012). Thus, parallel and high-performance computation will necessarily play a central role in attempts to derive insight from complex biological data. While the physics and chemistry communities have always been pioneers in high-performance scientific computing, there is a history of the use of such platforms for biological problems in areas such as phylogenetics (Korber et al. 2000), genetics (Conant et al. 2003), molecular dynamics and protein folding (Larson et al. 2002).

We will use three applications of next-generation sequencing as examples of the need not merely for "bigger" machines, hard drives and tools but for new approaches: shotgun metagenomics, population genomics and comparative genomics at the scale of the full tree of life.

The first, shotgun metagenomics through next-generation sequencing, offers the opportunity to probe almost any microbial ecosystem in effectively global detail. The price of this detail, however, is in large dataset sizes and long computational analyses: one of our own relatively modest surveys of 16 sheep microbiomes generated 134 GB of compressed sequence reads. Such samples can be analyzed in a matter of hours on a single CPU. However, emerging questions in metagenomics will require much larger data volumes and new analysis techniques:

• Analyses of the structure and function of the microbiome over time, either from birth or after diet change or antibiotic treatment (Looft et al. 2012; Yatsunenko et al. 2012).
• Comparative metagenomics across ecosystems to understand the rules structuring them.
• Expression analyses of the microbiome, exploring the enzyme expression patterns of the ecosystem, potentially also in a time-course framework (David et al. 2014).
• Tracing the spatial structure of a microbial ecosystem at either the landscape or global scale.

The second use of next-generation sequencing is population genomics: e.g., using sequencing to globally catalog all of the variation between hundreds to thousands of individuals from a population. Existing analyses are relatively computationally efficient because they rely on mapping reads from the sequenced individuals to a static reference genome (Langmead and Salzberg 2012). However, there are good reasons to believe that important genetic differences are missed with such tools because differences in gene copy number or genome architecture cannot be detected in this framework (Mackay et al. 2009). Hence, there will become a need to use much more computationally intensive genome assembly tools for analyzing these datasets (Pop 2009). Current tools and approaches are simply insufficient for genome assembly at these scales.

The third and final area using new sequencing paradigms is comparative genomics: better assembly tools would allow for genome-scale tree-of-life analyses. One can hardly imagine the insights that might be gleaned from having complete genomes for tens to hundreds of thousands of different species, but they would certainly revolutionize biology at all levels, from taxonomy to addressing the genotype-to-phenotype (G2P) problem (Pigliucci 2010).

These three areas of genomics also have a second commonality beyond their dependence on high-throughput sequences: they can be studied using a two-phase computational approach. The first phase involves string analysis problems involving matching and alignment operations on large sequence datasets. The second phase involves correlating the results of the first phase across genes and genomes to model or describe the resulting genomes and their function. Although there are a range of potential technologies that can be applied in the second phase, we will focus for brevity on the set of very effective approaches involving networks. In our work, we used an analysis of the metabolic networks of the microbes living in the vertebrate gut to show that the structure of this network differed in coherent ways depending on the animals' diet (Wolff et al. 2016). Again however, such network analyses will not easily scale to the larger dataset sizes that are coming, and there is a need for developing new tools that address this problem. In population genomics, similar problems are found: tools that rely on linear models of gene association between genes and traits can only explain a rather limited proportion of the genetically-linked phenotypic variation between individuals (Mackay et al. 2009). Instead, new network-aware tools for linking genome-scale variation data to phenotypes (Ayroles et al. 2009) are under development but will require new computational approaches to be fully tractable. Finally, analyses of complete genomes across the tree of life could help to address classic problems in cellular biology such as the structure of the gene regulatory network. A collaboratively hosted, configurable and extensible computational pipeline that provides high-performance pattern matching and network operations scaling to large datasets, as well as effective network visualization tools, would facilitate scientific discovery across all areas of genomics.

Question 2 Cyberinfrastructure Needed to Address the Research Challenge(s) (maximum ~1200 words): Describe any limitations or absence of existing cyberinfrastructure, and/or specific technical advancements in cyberinfrastructure (e.g. advanced computing, data infrastructure, software infrastructure, applications, networking, cybersecurity), that must be addressed to accomplish the identified research challenge(s).

All three applications above involve a string-matching phase, either mapping reads to a known database or the reads to each other (as in genome assembly). While many fast, special-purpose tools such as BowTie (Langmead and Salzberg 2012) exist for specific problems, these tools are not intended for the general analyses described above and are difficult to customize or port to arbitrary parallel architectures. For instance, we are not aware of generally-available tools for mapping metagenomic reads to known genes, still less one for the analysis of metagenomic metabolic networks. Instead, existing community tools, such as Galaxy and Qiime, are focused on taxa identification and hence canalize research in the area to a particular set of questions (Caporaso et al. 2010; Afgan et al. 2016). Likewise, existing network analysis tools like Cytoscape (Shannon et al. 2003) do not scale well beyond a few thousand nodes and limit the types of network analyses allowed.

However, existing work (including ours) suggests that both string matching and network analyses computations can be accelerated on massively parallel processors, such as Graphics Processing Units (GPUs) and Intel Phi devices. Originally designed as accelerators of image and video processing, GPUs are massively parallel devices that have been successfully used to accelerate a broad range of scientific computations from domains including computational chemistry, biology, physics, numerical analytics, weather prediction, computational finance, linear algebra and data mining (Nvidia Corporation 2016). Despite their widespread use, GPUs require specialized programming skills, and are particularly tricky to use for applications that include complex computation and memory access patterns, such as those considered here. Intel Phi coprocessors, which also offer massive parallelism but can be programmed using more traditional tools, are a viable alternative to GPUs, but still require specialized skills to achieve substantial speedup over CPU codes. Therefore, a genomics cyberinfrastructure that provides high-performance implementations of read matching and network analysis on massively parallel processors while hiding the complexity of these codes and the hardware from the scientists can enable discovery on large-scale datasets.

For read matching, many existing genomics tools are based on a combination of BLAST-like approaches (Altschul et al. 1997), local alignment (Smith and Waterman 1981) and suffix-tree methods. GPU implementations of these methods are available and can be integrated in the proposed computational pipeline. However, these methods have limitations in terms of scalability and kind of pattern matching operations they allow. Existing work has tackled scalability issues by either limiting the kind of pattern matching performed (for example, excluding inexact string matching), or by introducing filtering steps that limit the pattern matching operations to a subset of the inputs. These circumventions, however, limit the kinds of analyses that can be performed. More recently, there has been an increased interest in automata-based approaches to pattern matching. Efficient implementations of automata-based pattern matching have two potential advantages: first, they can scale to large datasets; second, they allow efficient implementations not only of exact-string matching, but also of more general forms of pattern matching (e.g., inexact matching with a given or variable number of insertions, deletions and substitutions), which can impact the specificity and sensitivity of the resulting read matches. The interest for automata-based solution has been also sparked by the proposal of high-performance implementations of automata-based pattern matching (for example, on FPGA, GPUs, and custom processors, like Micron's Automata Processor). In order to allow usability, a computational genomics cyber-infrastructure should hide from the user this complexity. In other words, the infrastructure should provide a high-performance pattern matching phase which: (i) is configurable in the kind of pattern matching operations to be performed and the matching sensitivity required, (ii) is accelerated on state-of-the-art parallel hardware, and (iii) seamlessly invokes the most suitable algorithmic implementation depending on the kind of pattern matching required and the scale of the input datasets.

Efficient implementations of network analysis tasks can enable more extensive and comprehensive explorations of the correlations between genes and genomes. The network analysis step can be based on several algorithms, such as the computation of the distance between all pairs of nodes in a network (Dijkstra 1959), the computation of connected components, and a variety of clustering and bipartite matching methods. Some of these algorithms have been successfully accelerated on GPUs (Harish and Narayanan 2007; Merrill et al. 2012; Li and Becchi 2013; Nasre et al. 2013; Shuai et al. 2013; Li et al. 2014; Li et al. 2015), particularly on large networks, and these implementations can be integrated in a computational genomics cyberinfrastructure. However, since the performance of network-based codes is very data-dependent, existing codes must be modified, and new codes need to be implemented, to be suitable to the particular scale and structure of networks emerging from computational genomics studies. Again, in order to allow usability, a computational genomics cyber-infrastructure should hide from the user this complexity. In other words, the infrastructure should provide a high-performance network analysis phase which: (i) is configurable and allows different kinds of network analysis, (ii) is accelerated on state-of-the-art parallel hardware, (iii) seamlessly invokes the most suitable implementation depending on the characteristics of the network to be analyzed.

We argue that the NSF CyVerse cyberinfrastructure system is a very nature place to deploy such new tools and resources for three reasons. First, considerable investment has already been made in this system. Second, the system allows for users with a range of abilities, from those using predefined analysis pipelines run through the Discovery Environment to those employing more individual approaches through tools like Atmosphere. Third and finally, CyVerse is already linked to large-scale parallel computing resources (e.g., Texas Advanced Computing Center) which have integrated coprocessors (GPUs and Intel Phi devices), allowing all users to access powerful parallel computing hardware.

Question 3 Other considerations (maximum ~1200 words, optional): Any other relevant aspects, such as organization, process, learning and workforce development, access, and sustainability, that need to be addressed; or any other issues that NSF should consider.

The origins of high performance computing were arguably in the realms of the physical sciences and applied mathematics and focused heavily on numerical approaches. As the need for high-performance computing in biology grows, the computing community will need to provide parallel tools for other classes of computations, such as those involving strings and networks. At the same time, to be cost-effective and sustainable, the solutions developed will need to run on community computing technologies such as mass-market CPUs, GPUs and other parallel processors. Providing such solutions within the constraints of cost and hardware availability requires tightly collaborative groups of computer scientists and biologists. These efforts, in turn, will contribute to the development of a workforce of biologists with computational skills (including parallel and high-performance computing), and of computer scientists with a better awareness of issues involved in real-world problems from biology.

REFERENCES:

Afgan E, Baker D, van den Beek M, Blankenberg D, Bouvier D, Cech M, Chilton J, Clements D, Coraor N, Eberhard C et al. 2016. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. Nucleic Acids Res 44: W3-W10.
Altschul SF, Madden TL, Schaffer AA, Zhang JH, Zhang Z, Miller W, Lipman DJ. 1997. Gapped Blast and Psi-Blast : A new-generation of protein database search programs. Nucleic Acids Res 25: 3389-3402.
Ayroles JF, Carbone MA, Stone EA, Jordan KW, Lyman RF, Magwire MM, Rollmann SM, Duncan LH, Lawrence F, Anholt RR et al. 2009. Systems genetics of complex traits in Drosophila melanogaster. Nat Genet 41: 299-307.
Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Pena AG, Goodrich JK, Gordon JI. 2010. QIIME allows analysis of high-throughput community sequencing data. Nature methods 7: 335-336.
Conant GC, Plimpton SJ, Old W, Wagner A, Fain PR, Pacheco TR, Heffelfinger G. 2003. Parallel Genehunter: Implementation of a linkage analysis package for distributed-memory architectures. Journal of Parallel and Distributed Computing 63: 674-682.
David LA, Maurice CF, Carmody RN, Gootenberg DB, Button JE, Wolfe BE, Ling AV, Devlin AS, Varma Y, Fischbach MA. 2014. Diet rapidly and reproducibly alters the human gut microbiome. Nature 505: 559-563.
Dijkstra EW. 1959. A note on two problems in connexion with graphs. Numerische Mathematik 1: 269-271.
Harish P, Narayanan PJ. 2007. Accelerating large graph algorithms on the GPU using CUDA. In Proceedings of the 14th international conference on High performance computing, pp. 197-208. Springer-Verlag, Goa, India.
Korber B, Muldoon M, Theiler J, Gao F, Gupta R, Lapedes A, Hahn BH, Wolinsky S, Bhattacharya T. 2000. Timing the ancestor of the HIV-1 pandemic strains. Science 288.
Lamanna M. 2004. The LHC computing grid project at CERN. Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment 534: 1-6.
Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nature Methods 9: 357-359.
Larson SM, Snow CD, Shirts M. 2002. Folding@ Home and Genome@ Home: Using distributed computing to tackle previously intractable problems in computational biology.
Li D, Becchi M. 2013. Deploying Graph Algorithms on GPUs: an Adaptive Solution. In Proceedings of the 2013 IEEE International Symposium on Parallel & Distributed Processing. IEEE Computer Society.
Li D, Chakradhar S, Becchi M. 2014. GRapid: A compilation and runtime framework for rapid prototyping of graph applications on many-core processors. In 2014 20th IEEE International Conference on Parallel and Distributed Systems (ICPADS), doi:10.1109/PADSW.2014.7097806, pp. 174-182.
Li D, Wu H, Becchi M. 2015. Nested Parallelism on GPU: Exploring Parallelization Templates for Irregular Loops and Recursive Computations. In 2015 44th International Conference on Parallel Processing (ICPP), doi:10.1109/ICPP.2015.107, pp. 979-988.
Looft T, Johnson TA, Allen HK, Bayles DO, Alt DP, Stedtfeld RD, Sul WJ, Stedtfeld TM, Chai B, Cole JR. 2012. In-feed antibiotic effects on the swine intestinal microbiome. Proceedings of the National Academy of Sciences 109: 1691-1696.
Mackay TF, Stone EA, Ayroles JF. 2009. The genetics of quantitative traits: challenges and prospects. Nat Rev Genet 10: 565-577.
Meehl GA, Covey C, Taylor KE, Delworth T, Stouffer RJ, Latif M, McAvaney B, Mitchell JF. 2007. The WCRP CMIP3 multimodel dataset: A new era in climate change research. Bulletin of the American Meteorological Society 88: 1383-1394.
Merrill D, Garland M, Grimshaw A. 2012. Scalable GPU graph traversal. In Proceedings of the 17th ACM SIGPLAN symposium on Principles and Practice of Parallel Programming, doi:10.1145/2145816.2145832, pp. 117-128. ACM, New Orleans, Louisiana, USA.
Nasre R, Burtscher M, Pingali K. 2013. Atomic-free irregular computations on GPUs. In Proceedings of the 6th Workshop on General Purpose Processor Using Graphics Processing Units, doi:10.1145/2458523.2458533, pp. 96-107. ACM, Houston, Texas.
National Research Council (U.S.). Policy and Global Affairs. 2012. The New global ecosystem in advanced computing : implications for U.S. competitiveness and national security. National Academies Press, Washington, D.C.

Nvidia Corporation. 2016. GPU-Accelerated Applications.
Pigliucci M. 2010. Genotype-phenotype mapping and the end of the 'genes as blueprint' metaphor. Philos Trans R Soc Lond B Biol Sci 365: 557-566.
Pop M. 2009. Genome assembly reborn: recent computational challenges. Brief Bioinform 10: 354-366.
Schadt EE, Linderman MD, Sorenson J, Lee L, Nolan GP. 2010. Computational solutions to large-scale data management and analysis. Nat Rev Genet 11: 647-657.
Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13: 2498-2504.
Shendure J, Ji H. 2008. Next-generation DNA sequencing. Nature Biotechnology 26: 1135-1145.
Shuai C, Beckmann BM, Reinhardt SK, Skadron K. 2013. Pannotia: Understanding irregular GPGPU graph applications. In Workload Characterization (IISWC), 2013 IEEE International Symposium on, doi:10.1109/iiswc.2013.6704684, pp. 185-195.
Smith TF, Waterman MS. 1981. Identification of common molecular subsequences. Journal of Molecular Biology 147: 195-197.
Stoughton C, Lupton RH, Bernardi M, Blanton MR, Burles S, Castander FJ, Connolly A, Eisenstein DJ, Frieman JA, Hennessy G. 2002. Sloan digital sky survey: Early data release. The Astronomical Journal 123: 485.
Wolff SM, Ellison MJ, Hao Y, Cockrum RR, Austin KJ, Baraboo M, Burch K, Lee HJ, Maurer T, Patil R et al. 2016. Diet shifts provoke complex and variable changes in the metabolic networks of the ruminal microbiome. Submitted.
Yatsunenko T, Rey FE, Manary MJ, Trehan I, Dominguez-Bello MG, Contreras M, Magris M, Hidalgo G, Baldassano RN, Anokhin AP. 2012. Human gut microbiome viewed across age and geography. Nature 486: 222-227.

Consent Statement

- “I hereby agree to give the National Science Foundation (NSF) the right to use this information for the purposes stated above and to display it on a publically available website, consistent with the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (https://creativecommons.org/licenses/by-nc-nd/4.0/legalcode).”